ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 95 (2005) 206–226

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

High breakdown estimators for principal components: the projection-pursuit approach revisited

Christophe Croux[a], Anne Ruiz-Gazen[b, c, *]

[a] *Department of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*

[b] *GREMAQ (U.M.R. CNRS 5604), University Toulouse I, Manufacture des Tabacs, 21, al. de Brienne 31042 Toulouse Cedex, France*

[c] *L.S.P. (U.M.R. CNRS 5583), University Toulouse III, 118 route de Narbonne, 31062 Toulouse Cedex, France*

Received 7 November 2003

Available online 18 September 2004

Abstract

Li and Chen (J. Amer. Statist. Assoc. 80 (1985) 759) proposed a method for principal components using projection-pursuit techniques. In classical principal components one searches for directions with maximal variance, and their approach consists of replacing this variance by a robust scale measure. Li and Chen showed that this estimator is consistent, qualitative robust and inherits the breakdown point of the robust scale estimator. We complete their study by deriving the influence function of the estimators for the eigenvectors, eigenvalues and the associated dispersion matrix. Corresponding Gaussian efficiencies are presented as well. Asymptotic normality of the estimators has been treated in a paper of Cui et al. (Biometrika 90 (2003) 953), complementing the results of this paper. Furthermore, a simple explicit version of the projection-pursuit based estimator is proposed and shown to be fast to compute, orthogonally equivariant, and having the maximal finite-sample breakdown point property. We will illustrate the method with a real data example.

AMS 1991 subject classification: 62F35; 62G35

Keywords: Breakdown point; Dispersion matrix; Influence function; Principal components analysis; Projection-pursuit; Robustness

* Corresponding author. University Toulouse III, 118 route de Narbonne, 31062 Toulouse Cedex, France. Fax: +33-05-61-22-55-63.

E-mail addresses: christophe.croux@econ.kuleuven.ac.be (C. Croux), ruiz@cict.fr (A. Ruiz-Gazen).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.08.002

1. Introduction

Classical principal components analysis (PCA) is very sensitive to outlying observations, since it is computed from eigenvectors and eigenvalues of the non-robust sample covariance or correlation matrix. Practitioners interpreting multivariate data solely on a classical PCA may therefore end up with wrong conclusions. This fact has been pointed out by many authors and has led to several robustifications of PCA (cf. [22, Chapter 10] for an overview). One may distinguish between two major types of approaches.

The first one calculates eigenvalues and eigenvectors based on a robust estimate of the covariance matrix. Originally, M-estimators for the covariance matrix were used for this (e.g. [13]). Their computation is not time consuming but they have a very low breakdown point in high dimensions. The breakdown point of an estimator measures the maximal percentage of the data points that may be contaminated before the estimate becomes completely corrupted and is very often used as a measure of robustness. Hence, high breakdown estimators for the covariance matrix are to be preferred. As such, the *minimum volume ellipsoid* estimator [29] was used by Naga and Antille [27]. The question of which robust covariance matrix estimator to use has recently been addressed by Croux and Haesbroeck [9]. They also computed influence functions and efficiencies for PCA based on robust estimators of the covariance or correlation matrix.

The second approach consists in calculating directly robust estimates of the eigenvalues and eigenvectors, without passing by a robust estimate of the covariance matrix. A projection-pursuit (PP) based method has been developed by Li and Chen [23] and was already mentioned by Huber [21]. Like classical PCA, they search for directions with maximal dispersion of the data projected on it. But instead of using the variance as a measure of dispersion, they use a robust scale estimator S_n as *projection-pursuit index*. For a sequence of observations $x_1, \ldots, x_n \in \mathbb{R}^p$, the first "eigenvector" is defined as

$$v_{S_n,1} = \underset{\|a\|=1}{\operatorname{argmax}} S_n(a^t x_1, \ldots, a^t x_n). \tag{1.1}$$

The associated "eigenvalue" is then by definition $\lambda_{S_n,1} = S_n^2((v_{S_n,1})^t x_1, \ldots, (v_{S_n,1})^t x_n)$. Suppose now that the first $k-1$ eigenvectors have already been found ($k > 1$). Then the kth eigenvector is defined as

$$v_{S_n,k} = \underset{\|a\|=1, a \perp v_{S_n,1}, \ldots, a \perp v_{S_n,(k-1)}}{\operatorname{argmax}} S_n(a^t x_1, \ldots, a^t x_n), \tag{1.2}$$

while the kth eigenvalue is defined as

$$\lambda_{S_n,k} = S_n^2((v_{S_n,k})^t x_1, \ldots, (v_{S_n,k})^t x_n). \tag{1.3}$$

Principal components scores are then given by the projections of the observations on the eigenvectors. Li and Chen [23] showed that the estimates inherit the breakdown point of the scale estimator S_n and are qualitative robust. As a by-product, a robust covariance estimate

can be deduced from the spectral decomposition:

$$C_{S_n} = \sum_{k=1}^{p} \lambda_{S_n,k} v_{S_n,k} v_{S_n,k}^t. \tag{1.4}$$

As was proven by Li and Chen, C_{S_n} is equivariant at elliptical models and consistent.

Li and Chen proposed to work with an M-estimator of scale for S_n, and applied a general PP algorithm for maximizing (1.2), leading to an iterative and complicated computer intensive method. This made their method quite unattractive to use in practice, in spite of the good theoretical properties. Nowadays, thanks to increasing computer power, there is a renewed interest in the PP approach to PCA. Filzmoser [14] applied it to a geostatistical problem, Boente et al. [4] in the context of common principal components, and Gather et al. [15] for robust sliced inverse regression.

After introducing the PP functionals in Section 2, we complete the theoretical study of Li and Chen by deriving the influence functions of the estimators of the eigenvalues, eigenvectors and the associated dispersion matrix (Section 3) and computing asymptotic variances (Section 4). In Sections 3 and 4, the influence function approach to robust statistics of Hampel et al. [16] is pursued. A formal treatment of the asymptotic distribution of the estimators (1.2) and (1.3) is presented in recent work of Cui et al. [11], hereby complementing the results of this paper. In Section 5, we propose a simple and explicit version of the PP-estimator. This estimator approximates $v_{S_n,k}$ and $\lambda_{S_n,k}$ by an easy to implement and fast algorithm, while remaining orthogonally equivariant and having a high finite-sample breakdown point. An application of this estimator to a real data set is presented in Section 6. Finally, Section 7 contains some conclusions.

2. The PP functionals

In order to derive the influence function, we first need to define the functionals of interest. Let G be an arbitrary p-dimensional distribution. Denote Ω_{p-1} the collection of all unit vectors in $\mathbb{R}^p$. For each $a \in \Omega_{p-1}$, denote G^a the distribution of $a^t X$ where $X \sim G$. Let S be an equivariant scale functional:

$$S(cY + b) = |c|S(Y) \tag{2.1}$$

for all real numbers c and b. By convention $T(Z) \equiv T(F)$, whenever $Z \sim F$ and for any statistical functional T.

We define the first population eigenvector $v_{S,1}(G)$ as the vector maximizing $S(G^a)$. The kth eigenvector $v_{S,k}(G)$ is defined by maximizing $S(G^a)$ over all $a \in \Omega_{p-1}$ subject to

$$v_{S,j}(G)^t a = 0$$

for all $j < k$. The robust eigenvalues of the distribution G are then given by

$$\lambda_{S,k}(G) = S^2(G^{v_{S,k}(G)}) \quad \text{for } k = 1, \ldots, p. \tag{2.2}$$

The associated robust dispersion matrix equals, using the spectral decomposition,

$$C_S(G) = \sum_{k=1}^{p} \lambda_{S,k}(G) v_{S,k}(G) v_{S,k}(G)^t. \tag{2.3}$$

Inserting for G the empirical distribution function yields the estimators $v_{S_n,k}$, $\lambda_{S_n,k}$ and C_{S_n} defined in the previous section.

The above-defined functionals $v_{S,k}$, $\lambda_{S,k}$, and C_S are orthogonally equivariant in the sense that

$$\begin{aligned} v_{S,k}(\Gamma X + b) &= \Gamma v_{S,k}(X), \quad \lambda_{S,k}(\Gamma X + b) = \lambda_{S,k}(X), \\ \text{and } C_S(\Gamma X + b) &= \Gamma C_S(X) \Gamma^t \end{aligned} \tag{2.4}$$

for every orthogonal matrix Γ and any translation vector $b \in \mathbb{R}^p$, with $1 \leqslant k \leqslant p$. In the context of PCA, orthogonal equivariance is sufficient, since even the classical procedures are only orthogonal equivariant. Affine equivariance of C_S is valid at an asymptotic level, within elliptical families.

Suppose that our observations $x_1, \ldots, x_n$ come from an elliptically symmetric model distribution H with location parameter μ and scatter matrix Σ. This means that the density of H can be written as

$$h(x) = \det(\Sigma)^{-1/2} g((x-\mu)^t \Sigma^{-1} (x-\mu)),$$

where $g: [0, \infty[\to \mathbb{R}^+$ is continuous. Furthermore, denote by $v_1, \ldots, v_p$ the eigenvectors of Σ and by $\lambda_1, \ldots, \lambda_p$ the corresponding eigenvalues, which we assume to verify $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.

The following lemma states that the projected distributions H^a belong to the same location-scale family. The proof results immediately from arguments given by Li and Chen [23, p. 760].

Lemma 1. *Let H be an elliptically symmetric distribution with parameters μ and Σ. Then there exists a univariate symmetric distribution F_0 such that*

$$H^a(y) = F_0\left(\frac{y - \mu^t a}{\sqrt{a^t \Sigma a}}\right). \tag{2.5}$$

The density of F_0 is given by $f_0(y) = \int \cdots \int g(y^2 + x_2^2 + \cdots + x_p^2)\, dx_2 \ldots dx_p$.

We will suppose that $S(F_0) = 1$, which can always be achieved by correction with a suitable consistency factor. By (2.1), it follows then that

$$S^2(H^a) = a^t \Sigma a, \tag{2.6}$$

which is a simple quadratic function. Maximization of (2.6) under the constraints stated above is easily done by using Lagrange multipliers. It is well known that the solutions $v_{S,k}(H)$ are nothing else but the eigenvectors of the matrix Σ and the corresponding $\lambda_{S,k}(H)$ are then the eigenvalues of Σ (in decreasing order). Therefore, $v_{S,k}(H) = v_k$ and $\lambda_{S,k}(H) = \lambda_k$ for $k = 1, \ldots, p$, while by (2.3) $C_S(H) = \Sigma$ implying Fisher consistency of the considered functionals at elliptically symmetric distributions.

The most important example is H multivariate normal, in which case $F_0 = \Phi$ the standard normal distribution. This is also the only situation in which the orthogonality of the eigenvectors implies independency of the different principal components.

The eigenvector and eigenvalue functionals are completely determined by the scale estimator S. Many robust scale estimators have been proposed in the literature, and we will focus on three of them. Perhaps the most well-known robust dispersion measure is the *median absolute deviation* (MAD). For a sample $\{y_1, \ldots, y_n\} \subset \mathbb{R}$ it is defined as

$$\mathrm{MAD}_n(y_1, \ldots, y_n) = 1.486 \operatorname*{med}_i |y_i - \operatorname*{med}_j y_j|,$$

where the constant 1.486 ensures consistency at normal distributions, i.e. $\mathrm{MAD}(\Phi) = 1$. The MAD has a 50% breakdown point, but a non-smooth influence function. M-estimators of scale can be seen as smooth versions of the MAD. Take an even function ρ, increasing for positive arguments, with $\rho(0) = 0$, then an M-estimator is defined as the solution of the following equation in s:

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathrm{med}_j\, y_j}{s}\right) = b$$

with $b = E_\Phi[\rho(Y)]$ to ensure consistency at normal distributions. If $\rho(\infty) = b/2$ the M-estimator has a 50% breakdown point [20, p. 110] and by choosing ρ properly it is possible to combine this with an arbitrarily high efficiency [6]. Another alternative to the MAD is the estimator Q_n of Rousseeuw and Croux [30], which is highly robust, fairly efficient and has an explicit definition since it is the first quartile of the pairwise differences between the data

$$Q_n(y_1, \ldots, y_n) = 2.2219\{|y_i - y_j|; 1 \leqslant i < j \leqslant n\}_{\binom{[n/2]+1}{2}:\binom{n}{2}}.$$

Again, we have for the associated functional $Q(\Phi)$=1.

3. Influence function

The influence function of a functional T at the distribution H is defined by

$$\mathrm{IF}(x; T, H) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)H + \varepsilon \Delta_x) - T(H)}{\varepsilon}, \tag{3.1}$$

where Δ_x has all its mass in x. It is a measure for the influence on the estimator T of an infinitesimal amount of contamination at x [16, Chapter 2]. We will suppose without any loss of generality that the location parameter of H equals zero, since the considered functionals are translation invariant. The proof of Theorem 1 is in the Appendix.

Theorem 1. *Let H be an elliptically symmetric distribution with $\mu = 0$ and scatter matrix Σ having distinct eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ with corresponding eigenvectors $v_1, \ldots, v_p$. Define F_0 as in Lemma 1. Assume that the function $(\varepsilon, y) \mapsto S((1-\varepsilon)F_0 + \varepsilon \Delta_y)$ is twice continuously differentiable at all points $(0, y)$. In particular, $\mathrm{IF}(y; S, F_0)$ needs to*

be differentiable and its derivative will be denoted by $\mathrm{IF}_1(y; S, F_0)$. *Then we obtain for the influence function of the eigenvalues*

$$\mathrm{IF}(x; \lambda_{S,k}, H) = 2\,\lambda_k\,\mathrm{IF}(\frac{x^t v_k}{\sqrt{\lambda_k}}; S, F_0) \tag{3.2}$$

and for the influence function of the eigenvectors

$$\begin{aligned}\mathrm{IF}(x; v_{S,k}, H) = &\sum_{j=1}^{k-1} \frac{\sqrt{\lambda_j}}{\lambda_k - \lambda_j}\mathrm{IF}_1(\frac{x^t v_j}{\sqrt{\lambda_j}}; S, F_0)(x^t v_k)v_j \\ &+ \sum_{j=k+1}^{p} \frac{\sqrt{\lambda_k}}{\lambda_k - \lambda_j}\mathrm{IF}_1(\frac{x^t v_k}{\sqrt{\lambda_k}}; S, F_0)(x^t v_j)v_j. \end{aligned}\tag{3.3}$$

By using a scale estimator with a bounded influence function, we obtain a bounded influence function for the eigenvalues. From (3.3) it also follows that the derivative of $\mathrm{IF}(y; S, F_0)$ determines the IF for the eigenvectors, and that scale estimators S having a smooth bounded derivative are to be preferred. However, the influence function for the eigenvectors may still become unbounded. Indeed, for most robust scale estimators $\mathrm{IF}_1(y; S, F_0)$ is bounded and even tends to or becomes 0 for $|y|$ tending to ∞, but the term $x^t v_j$ can make the influence function to go beyond all bounds. In fact, the following happens: denote by $(x^1, \ldots, x^p)^t$ the coordinates of the point x in the eigenvectors basis, so $x^j = x^t v_j$, and consider a scale functional S having an influence function with bounded derivative redescending to zero. First note that large values of x^1 have bounded influence on the estimation of all eigenvectors. Closer inspection of (3.3) reveals further that a huge value for x^j, $j > 1$ has limited influence on the eigenvectors $v_{S,k}$ for $k > j$. However, for $k < j$, a huge value of x^j combined with a smaller value of x^k may still yield a huge influence on the eigenvectors $v_{S,k}$ and $v_{S,j}$.

As a special case, consider $S^2(F) = \mathrm{VAR}(F)$. Since $\mathrm{IF}(y; S, F_0) = (y^2 - 1)/2$, Theorem 1 yields

$$\mathrm{IF}(x; \lambda_{S,k}, H) = (x^t v_k)^2 - \lambda_k$$

and

$$\mathrm{IF}(x; v_{S,k}, H) = \sum_{\substack{j=1 \\ j\neq k}}^{p} \frac{1}{\lambda_k - \lambda_j}\,(x^t v_k)(x^t v_j)v_j$$

for $k = 1, \ldots, p$. The above formulas for the classical estimator were already known and obtained by Critchley [5]. In Figs. 1 and 2 the IF for $\lambda_{S,1}$ and $v_{S,1}$ at a bivariate normal distribution $H = \mathrm{N}_2(0, \mathrm{diag}(2, 1))$ are pictured, once for S equal to the Q dispersion measure and once for the classical estimator. One observes that the shape of the influence function for the Q-based estimator is comparable to the classical estimator at the center of the distribution. Observations far away from the center of the distribution have a much smaller influence by using the Q estimator. The pictures confirm the boundedness of $\mathrm{IF}(x; \lambda_{Q,1}, H)$. For the eigenvectors, $\mathrm{IF}(x; v_{Q,1}, H)$ can still attain huge values, but only for smaller values of x^1 combined with huge values of x^2.

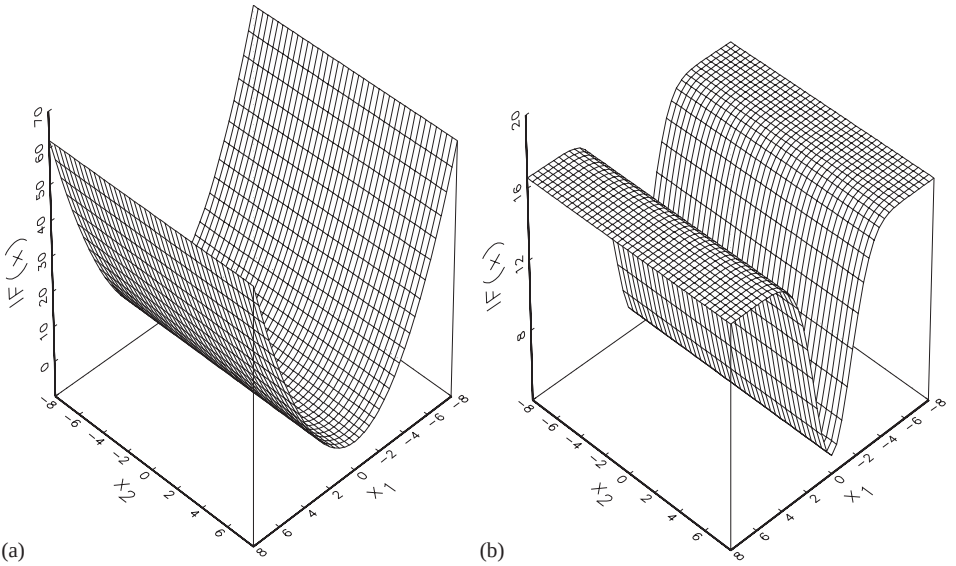

Fig. 1. Influence function of the largest eigenvalue for (a) the classical estimator and (b) the PP-estimator based on the Q dispersion measure, at $H = \mathrm{N}_2(0, \mathrm{diag}(2, 1))$.

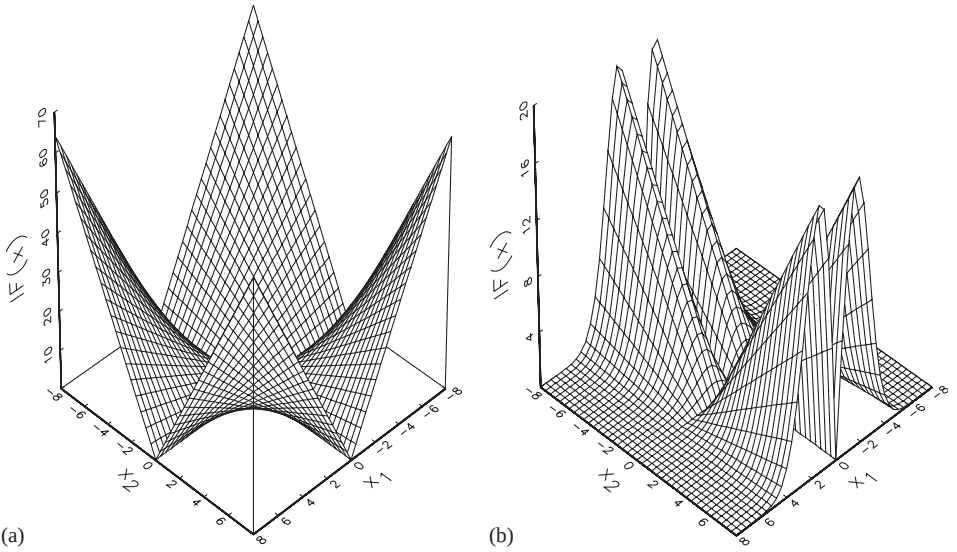

Fig. 2. Norm of the influence function of the eigenvector corresponding to the largest eigenvalue for (a) the classical estimator and (b) the PP-estimator based on the Q dispersion measure, at $H = \mathrm{N}_2(0, \mathrm{diag}(2, 1))$.

Using definition (2.3) and expressions (3.2) and (3.3), the influence function for the dispersion matrix functional C_S follows almost immediately:

$$\mathrm{IF}(x; C_S, H) = 2\sum_{k=1}^{p} \lambda_k \mathrm{IF}(\frac{x^t v_k}{\sqrt{\lambda_k}}; S, F_0) v_k v_k^t + \sum_{k=2}^{p}\sum_{j=1}^{k-1} \sqrt{\lambda_j}\mathrm{IF}_1(\frac{x^t v_j}{\sqrt{\lambda_j}}; S, F_0)(x^t v_k)[v_j v_k^t + v_k v_j^t]. \quad (3.4)$$

Once again, it appears clearly that a contaminant vector x may have an unbounded influence on C_S via the terms in $x^t v_k$ for $k = 2, \ldots, p$. It is amusing to notice that the complicated formula (3.4) reduces to $\mathrm{IF}(x; C_S, H) = xx^t - \Sigma$ by taking for S^2 the variance.

4. Asymptotic variances

As a corollary of Theorem 1, asymptotic variances for the estimators of eigenvectors and eigenvalues at the distribution H may be obtained using the heuristic formulas $\mathrm{ASV}(v_{S,k}; H) = E[\mathrm{IF}(X; v_{S,k}, H)\mathrm{IF}(X; v_{S,k}, H)^t]$ and $\mathrm{ASV}(\lambda_{S,k}; H) = E[\mathrm{IF}(X; \lambda_{S,k}, H)^2]$ (cf. [16, p. 92]). A rigorous proof of the asymptotic normality of the estimators can be found in Cui et al. [11]. Also Zhang [36] and Cui [10] studied the asymptotics of these estimators but in less general setting.

In case of a Gaussian model distribution, the asymptotic variances turn out to be particularly simple (cf. Appendix for a proof).

Corollary 1. *For $H = N(\mu, \Sigma)$ a multivariate normal distribution, asymptotic variances of the estimators of the eigenvectors and eigenvalues of Σ are given by*

$$\mathrm{ASV}(\lambda_{S,k}; H) = 4\lambda_k^2 \mathrm{ASV}(S, \Phi) \quad (4.1)$$

and

$$\mathrm{ASV}(v_{S,k}; H) = \sum_{\substack{j=1 \\ j\neq k}}^{p} \frac{\lambda_k \lambda_j}{(\lambda_j - \lambda_k)^2} v_j v_j{}^t E_\Phi[\mathrm{IF}_1(Y; S; \Phi)^2]. \quad (4.2)$$

As can be seen from (4.2), eigenvalues close to each other lead to high asymptotic variances of the eigenvector estimators, independently of the chosen estimator. Furthermore, projection indices S having an IF with exploding derivatives lead to exploding asymptotic variances.

To compare the precision of the different estimators, asymptotic efficiencies at normal distributions are computed as relative asymptotic variances with respect to the maximum likelihood estimator. The latter estimator is nothing else but the classical estimator [22, p. 50], which uses the standard deviation $\mathrm{STD} = \sqrt{\mathrm{VAR}}$ as projection index S. Define then

$$\mathrm{Eff}(\lambda_{S,k}; H) = \frac{\mathrm{ASV}(\lambda_{\mathrm{STD},k}; H)}{\mathrm{ASV}(\lambda_{S,k}; H)} = \frac{1}{2\mathrm{ASV}(S, \Phi)}$$

Table 1
Gaussian efficiencies for PP-based estimators of the eigenvalues and eigenvectors

S	Gaussian efficiency	Breakdown point
$\lambda_{S,k}$		
STD	100	0
MAD	36	50
M	$\leqslant 100$	50
Q	82	50
$v_{S,k}$		
STD	100	0
MAD	0	50
M	$\leqslant 33$	50
Q	67	50

The standard deviation (STD), the MAD, Q, and maximal breakdown M-estimators are considered as PP-indices.

and

$$\mathrm{Eff}(v_{S,k};\, H) = \frac{\mathrm{trace}[\mathrm{ASV}(v_{\mathrm{STD},k};\, H)]}{\mathrm{trace}[\mathrm{ASV}(v_{S,k};\, H)]} = \frac{1}{E_{\Phi}[\mathrm{IF}_1(Y;\, S;\, \Phi)^2]}$$

for $k = 1, \ldots, p$. Using the expressions for the influence functions of the scale estimators, Gaussian efficiencies are computed and reported in Table 1 (together with the value of the breakdown point) for different dispersion measures S. Note that the efficiencies are independent of the dimension p, which is in contrast with the efficiencies of PCA-estimators based on an eigenvalue decomposition of a robust covariance matrix (cf. [9]). The latter estimators, however, have a bounded influence function.

First of all, the efficiency of the estimator $\lambda_{S,k}$ is the same as that of the corresponding scale estimator. More surprisingly, it appears that $\mathrm{Eff}(v_{S,k};\, H)$ is identical to the Gaussian efficiency of the regression estimator based on the minimization of the dispersion S of the residuals. As such the obtained efficiency for the PP-based on the Q estimator is the same as that of the *least quartile difference* regression estimator of Croux et al. [7] (for normal error distributions). Using an M-estimator of scale yields the same efficiency as S-estimators of regression, which are defined as the minimizers of M-estimators of scale based on the residuals [33]. The Gaussian efficiency of 50% breakdown S-estimators of regression was shown to be bounded above by 33% [18]. The regression analogue for the MAD-based procedure is the *least median of squares* [31]. This estimator is known to have a slower rate of convergence and can therefore be said to have a zero efficiency. Notice that the MAD does not meet the differentiability condition of Theorem 1, since it has a jump in its influence function. To conclude, it seems that using Q as PP-index is a reasonable choice, since this scale estimator combines good efficiency with a smooth and bounded IF and the maximal breakdown point property.

5. A simple PP-based estimator

The estimators defined in (1.2) involve a non-trivial maximization problem, which has been considered as a major disadvantage of the approach (e.g. [25, p. 87]). If we suppose

that the first $k-1$ eigenvalues are already known, one needs to maximize the function

$$a \to S_n(x_i^t a; 1 \leqslant i \leqslant n) \tag{5.1}$$

under the conditions $a^t a = 1$ and $P_k a = a$. Here P_k stands for projection on the orthogonal complement of the space spanned by the first $k-1$ eigenvectors, and in particular $P_1 = I$. In general it will not be possible to obtain the exact solution to the above maximization problem, and therefore one needs to ressort to an approximation. Below, we outline a fast and simple algorithm for approximating the PP-estimators.

5.1. Description of the algorithm

Let $X = \{x_1, \ldots, x_n\}$ be the sample and $\hat{\mu}_n(X)$ a location estimate computed from this sample. Let $1 \leqslant q \leqslant p$ be the desired number of components to be computed and choose a scale estimator S_n as projection index.

- For $k = 1$, set $x_i^1 = x_i - \hat{\mu}_n(X)$ for $i = 1, \ldots, n$. Define then

$$A_{n,1}(X) = \{\frac{x_i^1}{||x_i^1||}; 1 \leqslant i \leqslant n\}$$

and set

$$\hat{v}_{S_n,1} = \underset{a \in A_{n,1}(X)}{\operatorname{argmax}} S_n(a^t x_1^1, \ldots, a^t x_n^1).$$

Compute then the scores on the first component as $y_i^1 = \hat{v}_{S_n,1}^t x_i^1$ for $i = 1, \ldots, n$.

- For $k = 2, \ldots, q$, define recursively
  1. for $i = 1, \ldots, n$, $x_i^k = x_i^{k-1} - y_i^{k-1} \hat{v}_{S_n,k-1}$,
  2. the set $A_{n,k}(X) = \{\frac{x_i^k}{||x_i^k||}; 1 \leqslant i \leqslant n\}$,
  3. the estimated eigenvector $\hat{v}_{S_n,k} = \operatorname{argmax}_{a \in A_{n,k}(X)} S_n(a^t x_1^k, \ldots, a^t x_n^k)$,
  4. for $i = 1, \ldots, n$, $y_i^k = \hat{v}_{S_n,k}^t x_i^k$

  yielding approximations for the eigenvectors and for the vector of scores on the kth principal component $(y_1^k, \ldots, y_n^k)^t$.

Approximations $\hat{\lambda}_{S_n,k}$, for $k = 1, \ldots, q$, for the eigenvalues and for the covariance matrix $\hat{C}_{S_n}$ are then computed as before, following (1.3) and (1.4). Note that the algorithm outlined above makes no smoothness assumptions on the scale estimate S_n, is simple and fast, and requires only $O(n)$ computing space.

It is easy to check that

$$A_{n,k}(X) = \{\frac{P_k(x_i - \hat{\mu}_n(X))}{\| P_k(x_i - \hat{\mu}_n(X))\|}; 1 \leqslant i \leqslant n\}, \tag{5.2}$$

with $P_k = (I - \sum_{j=1}^{k-1} \hat{v}_{S_n,j} \hat{v}_{S_n,j}^t)$. Hence

$$\hat{v}_{S_n,k} = \underset{a \in A_{n,k}(X)}{\operatorname{argmax}} S_n(a^t x_1, \ldots, a^t x_n), \tag{5.3}$$

so instead of scanning the whole space of possible solutions, as in (5.1), we will only check for vectors a belonging to the finite set $A_{n,k}$. In order to "work", the set $A_{n,k}$ in (5.3) should be quite dense in the region where the objective function reaches its maximum. Since the vectors belonging to $A_{n,k}$ point in the direction of the data, there is good hope that quite some of them will be close to the kth eigenvector, the latter one giving us the direction of maxima spread.

In [8] the approximations were directly computed as in (5.3). This, however, required explicit computation of the matrix P_k. The latter projection matrix has dimension $p \times p$, which may give rise to numerical problems for high-dimensional data-sets as was pointed out by Verboven et al. [34]. The recursive version of the algorithm outlined above does not suffer from this problem anymore.

As location estimator $\hat{\mu}_n$ we propose the spatial median or L_1-median. It is defined as

$$\hat{\mu}_n(X) = \operatorname*{argmin}_{\mu \in \mathbb{R}^p} \sum_{i=1}^{n} \|x_i - \mu\|, \tag{5.4}$$

where $\|\cdot\|$ stands for the Euclidean norm. This location estimator is orthogonally equivariant and has a 50% breakdown point. Its Gaussian efficiency is fairly high and increases with the dimension p. Since the objective function in (5.4) is convex, it can be computed extremely fast. Different algorithms for $\hat{\mu}_n(X)$ have been compared by Hössjer and Croux [19], and we chose to work with a gradient algorithm combined with stephalving. In case that software for computing $\hat{\mu}_n$ is not available to the user, a Matlab function can be retrieved from the homepage *http://www.econ.kuleuven.ac.be/christophe.croux*. Of course, also other robust location estimators can be taken here, but we recommend the L_1-median as it is sufficiently equivariant in the setting of PCA and fast to compute. Alternatively, the coordinatewise median could be taken as a crude approximation of $\hat{\mu}_n$.

The algorithm outlined above has been applied by Gather et al. [15] and Boente and Orellana [3] with satisfactory results. It is very easy to implement and the estimates are explicitly defined by (5.3). For example, when using the MAD as PP-index, the first eigenvalue estimate equals

$$\hat{\lambda}_{S_n,1} = \max_{1 \leqslant i \leqslant n} (\operatorname*{med}_j |y_i^t y_j - \operatorname*{med}_k y_i^t y_k|^2),$$

where $y_i = x_i - \hat{\mu}_n$. Another advantage of the procedure is that it allows for estimation of only the first q eigenvectors, without needing to compute all eigenvector estimates. In dimension reduction problems where p is huge and q is small, e.g. $q = 2$, this is an important feature of the procedure.

5.2. Some numerical experiments

To have an idea of the precision of the algorithm, a small simulation experiment was conducted. A sample of size n was generated from a p-variate normal distribution with mean zero and a diagonal covariance matrix with the elements $1, 2, \ldots, p$ on its diagonal. As a projection index the standard deviation was taken, since for STD is possible to find the exact maximum of (5.1). The value of $\hat{\lambda}_{S_n,1}$ was then computed by means of the approximative

Table 2
Precision of the simple PP-based estimator using STD as PP-index

p	5	10	20
$n = 50$	0.964	0.920	0.817
$n = 200$	0.985	0.940	0.851

Table 3
Computation time for computing the first 5 PCs using (a) the PP-based method with Q_n as index (b) the MCD estimator

p	5		10		20	
	PP	MCD	PP	MCD	PP	MCD
$n = 50$	0.06	1.82	0.06	3.18	0.06	6.32
$n = 200$	0.59	1.97	0.72	3.43	0.74	7.05

algorithm. As a measure of precision of the estimator, the fraction $\hat{\lambda}_{S_n,1}/\lambda_{S_n,1}$ is reported in Table 2 for various values of n and p. The results in table are averages over 10 simulation runs, with a standard error of at most 0.01 around the reported numbers.

From Table 2 one sees that the precision increases with n, which is no surprise since the number of search directions the algorithm considers increases with n. Furthermore, there is a loss in precision when the dimension p increases. In higher dimensions, there are much more trial directions needed to fill up the search space, yielding to a loss in precision when n is kept fixed.

To give an idea of the speed of the algorithm, computation times for obtaining the first 5 principal components using the simple algorithm and the Q_n estimator as PP-index are measured. They are compared with the time needed for carrying out a PCA using the eigenvalues and eigenvectors of a robust estimate of the covariance matrix. As robust estimate, the minimum covariance determinant (MCD) estimator was taken and computed with the FAST-MCD algorithm of Rousseeuw and Van Driessen [32]. The simulation scheme was the same as above, and MATLAB implementation of both methods were used. In Table 3, the average computation times over 10 simulation runs are reported (in seconds). Standard errors are negligible here.

From Table 3, it clearly follows that the simple PP-based estimator is indeed fast to compute. In all sampling schemes considered here the computation time was less than a second, using a 1400 MHz Pentium computer. Note that computing the first 5 components is often sufficient in applications. The gain in computation time w.r.t. the MCD procedure is important and increases further with p. A robust covariance matrix approach estimates all eigenvectors simultaneously, while the simple PP estimator takes advantage of the stepwise computation scheme and stops after having obtained the first 5 components. Furthermore, it is observed from Table 3 that the computation time of the PP-based method increases with the number of search directions in $A_{n,k}$, here equal to the sample size n.

5.3. Some properties of the simple PP-based estimators

The $\hat{v}_{S_n,k}$, $\hat{\lambda}_{S_n,k}$ and $\hat{\mathrm{C}}_{S_n}$ may not only be seen as approximations of (1.2), (1.3) and (1.4) but can be considered as estimators in their own right. They maintain the orthogonal equivariance property (2.4), since it is easy to verify that $A_{n,k}(\Gamma X) = \Gamma A_{n,k}(X)$, for each

$k = 1, \ldots, p$. Moreover, the estimator $\hat{C}_{S_n}$ can have a maximal finite sample breakdown point. Recall that the finite sample breakdown point of a scatter matrix estimator C_n is defined as

$$\varepsilon^*(C_n, X) = \min\{\frac{m}{n};\ \sup_{X'} \gamma(\hat{C}_n(X')) = \infty\},$$

where X' is obtained by replacing any m observations of X by arbitrary values and γ is the condition number of the indicated matrix (that is the largest divided by the smallest eigenvalue). We will make this explicit for the MAD PP index, but similar arguments can be given for Q_n or M-estimators of scale. First, we will slightly adapt the definition of the MAD. Instead of the median of deviations from the median, the $h_p = [(n + p + 1)/2]$ smallest deviation from the median will be taken:

$$\text{MAD}^*(y_1, \ldots, y_n) = 1.486\{|y_i - \underset{j}{\text{med}}\, y_j|;\ 1 \leqslant i \leqslant n\}_{h_p:n}. \tag{5.5}$$

The proof of the next proposition is in the Appendix.

Proposition 1. *Let S_n be the* MAD* *dispersion measure defined in* (5.5). *For every sample $X \subset \mathbb{R}^p$ in general position* (*meaning that no $p + 1$ points of X belong to the same hyperplane*) *we have*

$$\varepsilon^*(\hat{C}_{S_n}, X) \geqslant \frac{[(n - p + 1)/2]}{n}.$$

Note that it was proven by Davies [12] that $\varepsilon^*(C_n, X) \leqslant [n - p + 1/2]/n$ for any affine equivariant dispersion matrix estimator, but $\hat{C}_{S_n}$ is only orthogonal equivariant.

The real data example of the next section and the artificial data examples given in [8] show that the simple PP-estimators are well suited for exploratory data analysis, where emphasis is on finding the principal structure in the data. When focus is on statistical inference, this estimator may be too "simple", especially for smaller sample sizes. Indeed, simulation experiments have shown that while $\hat{v}_{S_n,k}$ and $\hat{\lambda}_{S_n,k}$ work fine for the first few principal components, they lack precision for estimation of the higher-order principal components. If one is also interested in accurate estimation of the higher-order principal components, then more sophisticated algorithms should be used. Xie et al. [35] experimented with simulated annealing techniques for optimizing (1.2). Ammann [1] used an iterative robust regression scheme for estimating the eigenvectors in reverse order, but this is restricted to a specific class of PP-indices. Also general purpose maximization routines could be used; if the objective function is differentiable, Newton-steps can be carried out. The simple PP-based estimator can then serve as a starting value.

6. Example

The McDonald and Schwing data set consists of $p = 16$ socioeconomic and climatological variables measured at each of the $n = 60$ Metropolitan Statistical Areas in the United States (see [26] for a description). The PP approach based on the Q_n scale measure, a

classical PCA, and a PCA based on the MCD estimator (with 50% breakdown point) of the covariance matrix have been applied. The simple version of the estimator, being described in the previous section, was implemented. Since the variables were measured on different scales, we first divided them by scale estimates. For the classical estimator this is equivalent with performing an eigenvalue analysis of the correlation matrix. We illustrate the use of PCA in an exploratory context by giving two graphical representations for each analysis: the projection of the observations on the first two principal axes and parallel boxplots of all p principal components. The first graphic is expected to display the main structure of the data set while the second picture can be used for outlier detection and the choice of the number of components to maintain (cf. [2]).

Fig. 3, first row, shows the pictures for a classical PCA. There still seems to be correlation present between the first and second principal component. This is explained by the outliers 29 and 48, whose influence makes the classical correlation coefficient equal to zero, while the big majority of the data still follows a linear pattern. Furthermore, outliers are mainly present on the first principal components, as can be seen from the sequence of boxplots. The reason is that outliers have attracted the first PC. By definition, the sample variance is maximal in this direction, but the spread of the big majority of the data isn't. We are not sure whether the first components really capture the main structure in the data, or are just reflecting the presence of outliers. It is risky to draw any conclusions from the outcomes, and we will indeed obtain different results with the robust method.

One could argue that outliers are made visible with the classical PCA approach, and that the analysis could be repeated after their deletion. But we do not know whether *all* outliers were detected on the first few components. (In fact, it this example it turned out by computing robust diagnostic measures that observation 18 is outlying, but this is not visible on the first two components of the classical PCA). It is even possible to construct artificial examples where all outliers will remain masked by a classical PCA. Moreover, often it is not clear cut to decide whether an observation is outlying or just at the "edge" of the data cloud formed by the good observations.

The second row of Fig. 3 represents the graphical displays for the PP method. Outliers are now present on almost all components, also on those of higher order, as can be seen from the sequence of boxplots. The plot of the first 2 PCs reveals no particular correlation structure, as it should be. No extreme outliers (like 29 and 48, confirmed by computing their robust Mahalanobis distances) are detected on the first and second component found by the PP approach, simply because they are not visible when projected in these directions for this example. In general, of course, it is quite possible that also the first few robust components reveal extreme outliers.

There is a group of data with higher values for the first principal component (observations 6-11-31-37). These observations correspond with areas in the south-east of the USA. The general principle is that the first robust principal components should provide the most interesting directions for the big majority of the data, independently of the position of possible outliers. One may say that robust PP PCA is a safe way to display the structure of the majority of the data on the first few principal axes, while the boxplots of the scores on the principal components allow to detect the principal axes which reveal outliers.

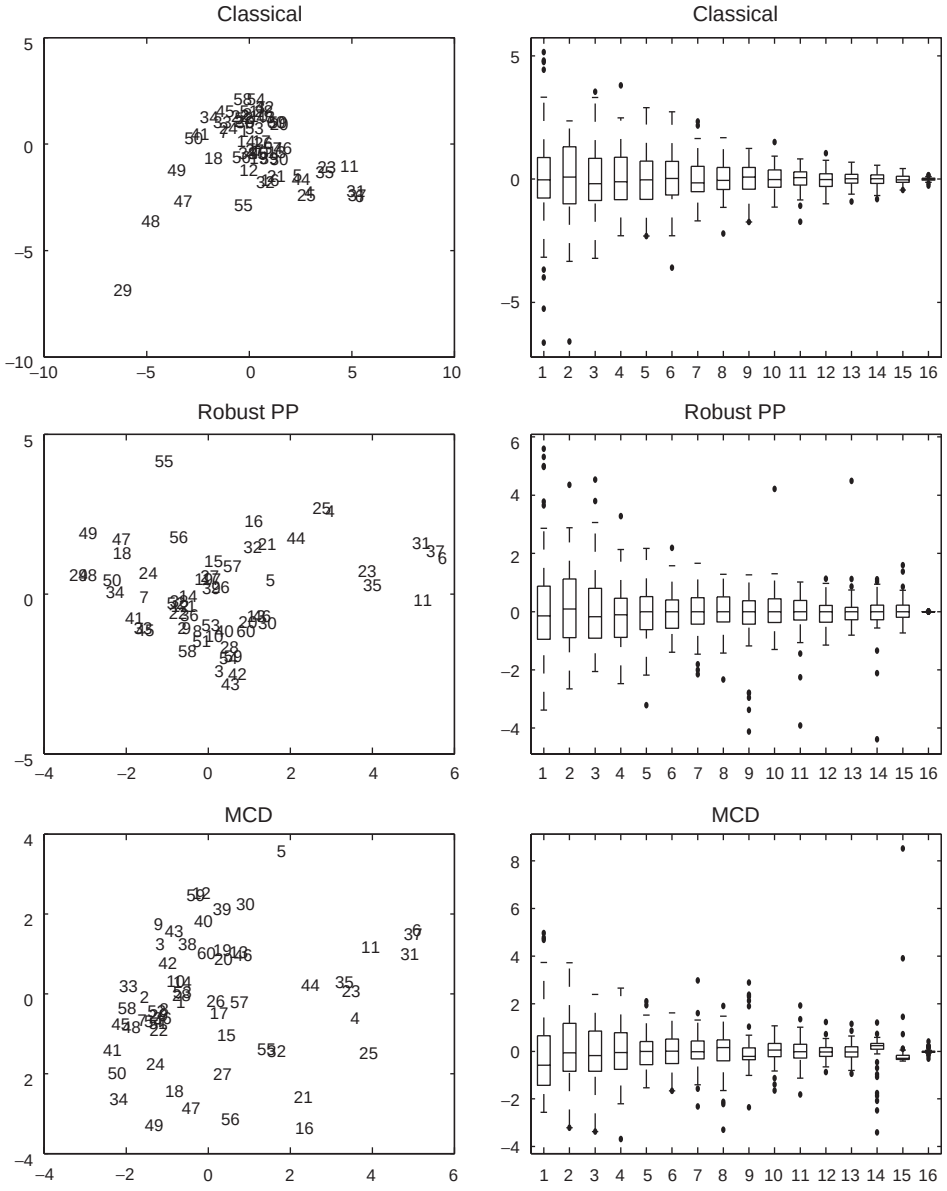

Fig. 3. Scatterplot of the scores on the first 2 principal components (left column) and boxplots of the scores on all principal components (right column) for the McDonald and Schwing data using (i) Classical PCA (ii) robust PP-based PCA (iii) PCA based on the robust MCD covariance matrix estimator.

Finally, note that the results for the robust covariance matrix approach (third row of Fig. 3) are similar to the PP-based analysis. Recall, however, that the latter method is much faster to compute.

7. Conclusion

In this paper, a general expression for the influence function of the PP-based principal components estimator has been derived. Using a robust scale as projection index, the IF for the eigenvalues is bounded, but this does not hold anymore for the eigenvectors. Hence, the estimators for the eigenvectors are sensible to small amounts of contaminations at some particular positions (see Fig. 2). The global robustness properties, however, may still be good, as can be verified by breakdown point calculations and computing maxbias curves of the PP-based eigenvector estimates (cf. [3] and unpublished work of [28]). Note that since eigenvectors live on the unit sphere, breakdown needs to be defined for estimates in a compact set (cf. [17]).

We also discussed a simple version of the PP-based estimator, which is easy to implement and fast to compute. We illustrated that this estimator is capable of retrieving the main structure of the majority of the data on the first principal axes. By looking at boxplots of the principal scores, outliers can be detected.

A major advantage of the PP-based approach is that the eigenvectors are found consecutively. In practice, it is often the case that one is only interested in the first 2 or 3 principal components. A lot of computation time can be saved by only searching for the first few eigenvectors, especially in high-dimensional settings.

The PP-based estimates can even be computed for data matrices with more variables than observations. The situation $n < p$ is frequently encountered in practical applications (e.g. [24]). Note that the robust PCA approach based on robust estimators of covariance is not applicable in this setting.

Appendix

Proof of Theorem 1. Let $\varepsilon > 0$, $x \in \mathbb{R}^p$, and consider the contaminated distribution $H_{\varepsilon,x} = (1-\varepsilon)H + \varepsilon\Delta_x$. Use the shorthand notations $v_{k,\varepsilon} = v_{S,k}(H_{\varepsilon,x})$ and $\lambda_{k,\varepsilon} = \lambda_{S,k}(H_{\varepsilon,x})$, for $k = 1, \ldots, p$. We fix k and want to compute $\mathrm{IF}(x; v_{S,k}, H) = \frac{\partial}{\partial\varepsilon} v_{k,\varepsilon}|_{\varepsilon=0}$ and $\mathrm{IF}(x; \lambda_{S,k}, H) = \frac{\partial}{\partial\varepsilon}\lambda_{k,\varepsilon}|_{\varepsilon=0}$.

Influence function for the eigenvectors: The vector $v_{k,\varepsilon}$ is maximizing $S(H^a_{\varepsilon,x})$ under the constraints that $v^t_{k,\varepsilon} v_{k,\varepsilon} = 1$ and $v^t_{k,\varepsilon} v_{j,\varepsilon} = 0$ for $j = 1, \ldots, k-1$, resulting in the Lagrangian function

$$\mathcal{L}(a, \gamma, \alpha_1, \ldots, \alpha_{k-1}) = S^2(H^a_{\varepsilon,x}) - \gamma(a^t a - 1) - \sum_{j=1}^{k-1} \alpha_j a^t v_{j,\varepsilon}.$$

Since $v_{k,\varepsilon}$ maximizes this Lagrangian, it needs to verify

$$\psi(\varepsilon) = 2\gamma v_{k,\varepsilon} + \sum_{j=1}^{k-1} \alpha_j v_{j,\varepsilon} \tag{A.1}$$

with

$$\psi(\varepsilon) = \frac{\partial}{\partial a} S^2(H^a_{\varepsilon,x})\big|_{a = v_{k,\varepsilon}}. \tag{A.2}$$

From the side restrictions on $v_{k,\varepsilon}$ and (A.1) it follows that $\psi(\varepsilon)^t v_{k,\varepsilon} = 2\gamma$ and $\psi(\varepsilon)^t v_{j,\varepsilon} = \alpha_j$, for $j = 1, \ldots, k-1$. Eq. (A.1) can therefore be rewritten as

$$\psi(\varepsilon) = \sum_{j=1}^{k} (\psi(\varepsilon)^t v_{j,\varepsilon}) v_{j,\varepsilon}. \tag{A.3}$$

Derivation of (A.3) yields

$$\begin{aligned} \frac{\partial}{\partial \varepsilon}\psi(\varepsilon)\big|_{\varepsilon=0} = \sum_{j=1}^{k} \Bigg(& \psi(0)^t \mathrm{IF}(x; v_{S,j}, H) v_j + v_j^t \frac{\partial}{\partial \varepsilon}\psi(\varepsilon)\big|_{\varepsilon=0} v_j \\ & + \psi(0)^t v_j \mathrm{IF}(x; v_{S,j}, H) \Bigg). \end{aligned} \tag{A.4}$$

Now $\psi(0) = \frac{\partial}{\partial a} S^2(H^a)\big|_{v_k} = 2\Sigma v_k = 2\lambda_k v_k$ and therefore $\psi(0)^t v_j = 0$ for $j = 1, \ldots, k-1$. Denote now $P_{k+1} = I_p - \sum_{j=1}^{k} v_j v_j^t$, then (A.4) becomes

$$P_{k+1} \frac{\partial}{\partial \varepsilon}\psi(\varepsilon)\big|_{\varepsilon=0} = 2\lambda_k \sum_{j=1}^{k} (v_k^t \mathrm{IF}(x; v_{S,j}, H)) v_j + 2\lambda_k \mathrm{IF}(x; v_{S,k}, H). \tag{A.5}$$

On the other hand, using the chain derivation rule, we obtain from (A.2) that

$$\begin{aligned} \frac{\partial}{\partial \varepsilon}\psi(\varepsilon)\big|_{\varepsilon=0} &= \frac{\partial}{\partial a} \frac{\partial}{\partial a^t} S^2(H^a)\big|_{a=v_k} \mathrm{IF}(x; v_{S,k}, H) + \frac{\partial}{\partial \varepsilon} \frac{\partial}{\partial a} S^2(H^a_{\varepsilon,x})\big|_{a=v_k, \varepsilon=0} \\ &= 2\Sigma \mathrm{IF}(x; v_{S,k}, H) + \frac{\partial}{\partial a} \mathrm{IF}(a^t x; S^2, H^a)\big|_{a=v_k}. \end{aligned} \tag{A.6}$$

Using equivariance (2.1) of the scale estimator and Lemma 1 allows to compute the derivative in the above equation:

$$\begin{aligned} \frac{\partial}{\partial a} \mathrm{IF}(a^t x; S^2, H^a)\big|_{a=v_k} &= \frac{\partial}{\partial a} a^t \Sigma a \mathrm{IF}\left(\frac{a^t x}{\sqrt{a^t \Sigma a}}; S^2, F_0\right)\Big|_{a=v_k} \\ &= 2\lambda_k v_k \mathrm{IF}\left(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0\right) \\ &\quad + \lambda_k \mathrm{IF}_1\left(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0\right) \left[\frac{x}{\sqrt{\lambda_k}} - \frac{v_k^t x}{\sqrt{\lambda_k}} v_k\right]. \end{aligned}$$

Recall that IF_1 stands for the first derivative of the influence function of the scale functional S. Now it follows from (A.6) that Eq. (A.5) can be rewritten as

$$2P_{k+1}\Sigma\mathrm{IF}(x; v_{S,k}, H) + \lambda_k \mathrm{IF}_1\left(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0\right)\frac{P_{k+1}x}{\sqrt{\lambda_k}}$$
$$= 2\lambda_k \sum_{j=1}^{k} (v_k^t \mathrm{IF}(x; v_{S,j}, H))v_j + 2\lambda_k \mathrm{IF}(x; v_{S,k}, H)$$

or

$$(P_{k+1}\Sigma - \lambda_k I_p)\mathrm{IF}(x; v_{S,k}, H) = -\frac{\sqrt{\lambda_k}}{2}\mathrm{IF}_1\left(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0\right)P_{k+1}x$$
$$+\lambda_k \sum_{j=1}^{k} (v_k^t \mathrm{IF}(x; v_{S,j}, H))v_j. \tag{A.7}$$

Now $P_{k+1}\Sigma - \lambda_k I_p = \sum_{j=k+1}^{p} \lambda_j v_j v_j^t - \lambda_k I_p$ is a rank $p-1$ matrix with generalized inverse

$$(P_{k+1}\Sigma - \lambda_k I_p)^- = \sum_{j=k+1}^{p} \frac{1}{\lambda_j - \lambda_k} v_j v_j^t - \sum_{j=1}^{k-1} \frac{1}{\lambda_k} v_j v_j^t.$$

Since derivating $v_{k,\varepsilon}^t v_{k,\varepsilon} = 1$ implies that $\mathrm{IF}(x; v_{S,k}, H)^t v_k = 0$, $\mathrm{IF}(x; v_{S,k}, H)$ has no component in the direction of v_k and (A.7) determines $\mathrm{IF}(x; v_{S,k}, H)$ uniquely. We obtain

$$\mathrm{IF}(x; v_{S,k}, H) = -\frac{\sqrt{\lambda_k}}{2} \sum_{j=k+1}^{p} \frac{\mathrm{IF}_1\left(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0\right)}{\lambda_j - \lambda_k}(v_j^t x)v_j$$
$$-\sum_{j=1}^{k-1} (\mathrm{IF}(x; v_{S,j}, H)^t v_k)v_j \tag{A.8}$$

for every $k = 1, \ldots, p$. From this recursive relation we deduce that

$$\mathrm{IF}(x; v_{S,j}, H)^t v_k = \frac{-\sqrt{\lambda_j}}{\lambda_k - \lambda_j}\mathrm{IF}_1\left(\frac{x^t v_j}{\sqrt{\lambda_j}}; S^2, F_0\right)(v_k^t x)$$

for $j < k$, so that (A.8), combined with $\mathrm{IF}(y; S^2, F_0) = 2\,\mathrm{IF}(y; S, F_0)$, yields the result (3.3).

Influence function for the eigenvalues: By definition $\lambda_{k,\varepsilon} = S^2(H_{\varepsilon,x}^{v_{k,\varepsilon}})$, Application of the chain rule yields

$$\mathrm{IF}(x; \lambda_{S,k}, H) = \frac{\partial}{\partial\varepsilon} S^2(H_{\varepsilon,x}^{v_k})\Big|_{\varepsilon=0} + \left(\frac{\partial}{\partial a} S^2(H^a)\Big|_{a=v_k}\right)^t \mathrm{IF}(x; v_{S,k}, H)$$
$$= \mathrm{IF}(v_k^t x; S^2, H^{v_k}) + \psi(0)^t \mathrm{IF}(x; v_{S,k}, H)$$
$$= \lambda_k \mathrm{IF}\left(\frac{v_k^t x}{\sqrt{v_k^t \Sigma v_k}}; S^2, F_0\right)$$

$$= 2\lambda_k \mathrm{IF}(\frac{x^t v_k}{\sqrt{\lambda_k}}; S^2, F_0)$$

for $k = 1, \ldots, p$ which ends the proof. □

Proof of Corollary 1. Expression (4.1) for the eigenvalues is trivial. For the eigenvectors we need to compute $E[\mathrm{IF}_1(X; v_{S,k}, H)\mathrm{IF}_1(X; v_{S,k}, H)^t]$ where $\mathrm{IF}(X; v_{S,k}, H)$ is given by (3.3). Suppose w.l.o.g. that $\mu = 0$.

Note that, for a fixed $1 \leqslant k \leqslant p$,

$$E_H[\mathrm{IF}_1(\frac{X^t v_j}{\sqrt{\lambda_j}}; S, F_0)\mathrm{IF}_1(\frac{X^t v_l}{\sqrt{\lambda_l}}; S, F_0)(x^t v_k)^2] = 0$$

for $1 \leqslant j \neq l < k$. Therefore we use that $X^t v_l$, $X^t v_j$ and $X^t v_k$ are independent and $E_H[\mathrm{IF}_1(\frac{X^t v_j}{\sqrt{\lambda_j}}; S, F_0)] = E_\Phi[\mathrm{IF}'(Y; S, \Phi)] = 0$ since the influence function of a scale estimator at a symmetric distribution is symmetric. On the other hand

$$\begin{aligned} E_H[(\mathrm{IF}_1(\frac{x^t v_j}{\sqrt{\lambda_j}}; S^2, F_0))^2 (x^t v_k)^2] &= \lambda_k E_\Phi[\mathrm{IF}_1(Y; S, \Phi)^2] E_\Phi[Y^2] \\ &= \lambda_k E_\Phi[\mathrm{IF}_1(Y; S, \Phi)^2]. \end{aligned}$$

Furthermore, using similar arguments as above, we obtain

$$E_H[\mathrm{IF}_1(\frac{X^t v_k}{\sqrt{\lambda_k}}; S, F_0)^2 (X^t v_j)(X^t v_l)] = 0$$

for $k < j \neq l \leqslant p$, while

$$E_H[\mathrm{IF}_1(\frac{X^t v_k}{\sqrt{\lambda_k}}; S, F_0)^2 (X^t v_j)^2] = \lambda_j E_\Phi[\mathrm{IF}'(Y; S, \Phi)^2].$$

Finally, for the cross terms

$$E_H[\mathrm{IF}_1(\frac{X^t v_k}{\sqrt{\lambda_k}}; S, F_0)(X^t v_l)\mathrm{IF}_1(\frac{X^t v_j}{\sqrt{\lambda_j}}; S, F_0)(X^t v_k)] = 0$$

for $1 \leqslant j < k < l \leqslant p$. Using the above equalities, (4.2) is readily obtained. □

Proof of Proposition 1. Let $S_n = \mathrm{MAD}_n^*$ and take X a sample in general position. Define $\delta = \frac{1}{2} \inf\{\rho > 0 |$ there exists a hyperplane $\mathcal{H}$ such that at least $(p+1)$ points of X are within a distance ρ of $\mathcal{H}\}$, where distance stands for orthogonal euclidean distance. Since X is in general position, we have $\delta > 0$. Furthermore, let $M = \sup_i \|X_i\|$. Replace now $m \leqslant [(n-p-1)/2]$ points of X by arbitrary values, and denote X' the resulting contaminated sample. Of course, X' still contains $n - m$ observations from X. We will prove that

$$\gamma(\hat{\mathrm{C}}_{S_n}(X')) = \frac{\hat{\lambda}_{S_n,1}(X')}{\hat{\lambda}_{S_n,p}(X')} \leqslant \frac{2M}{\delta},$$

showing that no breakdown occurs and thus $\varepsilon^*(\hat{\mathrm{C}}_{S_n}, X) \geqslant [(n-p+1)/2]/n$.

First of all, note that for any $a \in A_{n,k}(X')$, we have $|a^t x_i'| \leqslant \|a\| \|x_i'\| \leqslant M$ at least $n - m = [(n+p)/2] + 1$ times. Therefore $|\operatorname{med}_j a^t x_j'| \leqslant M$ and $|a^t x_i' - \operatorname{med}_j a^t x_j'| \leqslant 2M$ at least $n - m \geqslant h_p$ times, so that $S_n(a^t X') \leqslant 2M$. Since this holds for any $a \in A_{n,k}(X')$, we have $\hat{\lambda}_{S_n,1}(X') \leqslant 2M$.

On the other hand, take $a \in \Omega_{p-1}$ orthogonal to the space spanned by the first $p-1$ eigenvectors $\hat{v}_{S_n,1}(X'), \ldots, \hat{v}_{S_n,p-1}(X')$. This vector a will then be equal to the last eigenvector of X. Consider the hyperplane $\mathcal{H}_a' = \{x \in \mathbb{R}^p | a^t x = \operatorname{med}_j (a^t x_j')\}$. By definition of δ we have that $|a^t x_i' - \operatorname{med}_j a^t x_j'| > \delta$ at least $n - m - p$ times. Since $h_p \geqslant n - (n - m - p) + 1 = [(n + p + 1)/2]$, this yields $\hat{\lambda}_{S_n,p}(X') = S_n(a^t X') > \delta$. □

Acknowledgements

We wish to thank the reviewers for helpful comments. This research has been supported by the Research Fund K.U. Leuven and the "Fonds voor Wetenschappelijk Onderzoek" (Contract number G.0385.03).

References

[1] L.P. Ammann, Robust singular value decompositions: a new approach to projection pursuit, J. Amer. Statist. Assoc. 88 (1993) 505–514.
[2] P.H. Besse, A. de Falguerolles, Application of resampling methods to the choice of dimension in principal components analysis, in: W. Härdle, L. Simar (Eds.), Computer Intensive Methods in Statistics, Physica-Verlag, Wurzburg, 1993.
[3] G. Boente, L. Orellana, A robust approach to common principal components, in: L.T. Fernholz, S. Morgenthaler, W. Stahel (Eds.), Statistics in Genetics and in the Environmental Sciences, Birkhauser, Basel, 2001, pp. 117–147.
[4] G. Boente, A.M. Pires, I.M. Rodrigues, Influence functions and outlier detection under the common principal components model: a robust approach, Biometrika 89 (2002) 861–875.
[5] F. Critchley, Influence in principal components analysis, Biometrika 72 (1985) 627–636.
[6] C. Croux, Efficient high-breakdown M-estimators of scale, Statist. Probab. Lett. 19 (1994) 371–379.
[7] C. Croux, P.J. Rousseeuw, O. Hössjer, Generalized S-estimators, J. Amer. Statist. Assoc. 89 (1994) 1271–1281.
[8] C. Croux, A. Ruiz-Gazen, A fast algorithm for robust principal components based on projection pursuit, in: A. Prat (Ed.), Compstat: Proceedings in Computational Statists, Physica-Verlag, Heidelberg, 1996, pp. 211–216.
[9] C. Croux, G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika 87 (2000) 603–618.
[10] H. Cui, A sufficient and necessary condition for a PP-estimator of scatter matrix to be quantitatively robust, Chinese J. Appl. Probab. Statist. 8 (1993) 113–121.
[11] H. Cui, X. He, K.W. Ng, Asymptotic distributions of principal components based on robust dispersions, Biometrika 90 (2003) 953–966.
[12] L. Davies, Asymptotic behavior of S-estimators of multivariate location estimators and dispersion matrices, Ann. Statist. 15 (1987) 1269–1292.
[13] S.J. Devlin, R. Gnanadesikan, J.R. Kettenring, Robust estimation of dispersion matrices and principal components, J. Amer. Statist. Assoc. 76 (1981) 354–362.
[14] P. Filzmoser, Robust principal component and factor analysis in the geostatistical treatment of environmental data, Environmetrics 10 (1999) 363–375.

[15] U. Gather, T. Hilker, C. Becker, A robustified version of sliced inverse regression procedures, in: L.T. Fernholz, S. Morgenthaler, W. Stahel (Eds.), Statistics in Genetics and in the Environmental Sciences, Birkhauser, Basel, 2001, pp. 147–157.
[16] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, New York, 1986.
[17] X. He, D.G. Simpson, Robust direction estimation, Ann. Statist. 20 (1992) 351–369.
[18] O. Hössjer, On the optimality of S-estimators, Statist. Probab. Lett. 14 (1992) 413–419.
[19] O. Hössjer, C. Croux, Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter, Nonparametric Statist. 4 (1995) 293–308.
[20] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[21] P.J. Huber, Projection pursuit, Ann. Statist. 13 (1985) 435–525.
[22] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer, New York, 2002.
[23] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, J. Amer. Statist. Assoc. 80 (1985) 759–766.
[24] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, Robust principal components for functional data, Test 8 (1999) 1–73.
[25] R.A. Maller, Some consistency results on projection pursuit estimators of location and scale, Canad. J. Statist. 17 (1989) 81–90.
[26] G.C. McDonald, R.C. Schwing, Instabilities of regression estimates relating air pollution to mortality, Technometrics 15 (1973) 463–481.
[27] R. Naga, G. Antille, Stability of robust and non-robust principal component analysis, Comp. Statist. Data Anal. 10 (1990) 169–174.
[28] Z. Patak, Robust principal components, Master Thesis, Department of Statistics, University of British Columbia, 1991.
[29] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, vol. B, Reidel, Dordrecht, 1985, pp. 283–297.
[30] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Amer. Statist. Assoc. 88 (1993) 1273–1283.
[31] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[32] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[33] P.J. Rousseeuw, V.J. Yohai, Robust regression by means of S-estimators, in: J. Franke, W. Härdle, R.D. Martin (Eds.), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics, vol. 26, Springer, New York, 1984.
[34] S. Verboven, P.J. Rousseeuw, M. Hubert, An improved algorithm for robust PCA, in: COMPSTAT: Proceedings in Computational Statistics,Springer, Berlin, 2000, pp. 481–486.
[35] Y. Xie, J. Wang, Y. Liang, L. Sun, X. Song, R. Yu, Robust principal components analysis by projection pursuit, J. Chemometrics 7 (1993) 527–541.
[36] J. Zhang, Asymptotic theories for the robust PP estimators of the principal components and dispersion matrix (III, bootstrap confidence sets bootstrap tests), System Sci. Math. Sci. 4 (1991) 289–301.